随机森林在高校信息碎片化整合中的应用*

张文德[1] 程涵[1] 刘田[2] 曾金晶[2]
[1] 福州大学信息管理研究所 福州 350108 [2] 福建农林大学图书馆 福州 350002

摘要：[目的/意义]面对高校信息呈现碎片化的趋势，提出高校信息碎片化整合流程，并应用随机森林算法构建高校信息碎片化整合的特征选择模型。[方法/过程]基于高校信息整合的发展现状与存在问题，分析随机森林算法原理及优势，将其运用到高校信息碎片化整合过程的特征选择模型中，并以高校贫困生认定为例，对该模型加以验证。[结果/结论]随机森林算法在高校信息整合特征选择上表现出较高的准确性和有效性，为高校信息碎片化整合提供了一种新的思路。

关键词：随机森林 碎片化 信息整合 特征选择
分类号：G203
DOI：10.13266/j.issn.0252-3116.2018.07.014

1 引言

20世纪90年代初，在政府部门的高度重视下我国高校信息化建设快速发展。纵观高校信息化建设的发展历程，可以将高校信息化整合过程划分为3个阶段：基于数据的整合阶段、基于信息的整合阶段和基于知识的整合阶段[1-3]。在高校信息化建设之初，为了满足单一部门的信息需求，建立了一个个相互独立的信息系统。随着"信息孤岛"的形成，高校异构信息系统的集成问题逐步受到重视。基于数据的整合这一阶段的主要目的是利用中间件及数据仓库等技术，通过消除系统的分布性和异构性，实现异构信息系统数据的统一存储。基于信息的整合是在数据集成的基础上发展而来的，这一阶段主要从业务需求的角度出发，通过企业架构等方式整合满足某一业务流程的信息资源，实现异构信息系统之间信息交流与共享。但随着云计算、移动互联网、物联网等技术的应用，高校信息系统数据呈指数增长，正逐步进入"大数据"时代，基于知识的整合阶段除了要实现对海量多源异构数据的共享外，还要充分挖掘其潜在价值，实现知识集成及创新，从而为用户管理决策提供支持与帮助。

目前，实现基于知识的信息整合主要是利用数据挖掘和机器学习算法。常桐善[4]认为数据挖掘技术能帮助大学管理人员更好地分析数据，从而获取潜藏的、有用的信息和知识，最终提高决策效率。廖凤露、周庆[5]采用朴素贝叶斯模型对学生的就业能力进行预测，为学生的就业工作提供帮助。施佺、钱源和孙玲[6]将关联规则算法和聚类算法等数据挖掘技术运用到网络学习的监管中，为了解网络学习效果、改进网络学习过程提供参考。舒忠梅、徐晓东[7]利用逐步回归和决策树分析等数据挖掘方法对大学生满意度进行分析，探究了影响学生满意度的因素，为高校人才培养提供了参考依据。何世明、沈军[8]利用BP神经网络和聚类分析技术，通过挖掘隐藏在数据中有用的规律和知识，提出了一个适合网络教学的学习评价方法，为教学评估提供决策支持。刘美玲、李熹和李永胜[9]提出了一种基于K-Means算法的成绩聚类分析方法，以说明数据挖掘技术在教育系统中的应用。

纵观目前在高校数据挖掘领域的所有研究，不管是决策树、神经网络还是关联、聚类算法都存在以下问题：①数据挖掘算法仅仅运用到部分教育问题，如教学评估、学习效率监督等，即算法模型仅能解决某一个或一类问题，并不具有很强的通用性和推广性；②现有的

* 本文系中国高教学会2016年度教育信息化专项课题"基于高校信息碎片化的信息整合构建研究"（项目编号：2016XXZD01）、2016年福建省教育厅中青年项目"基于不含有主键大数据精简的直觉模糊决策方法及其应用研究"（项目编号：JAT160097）和"基于中智集的高校信息系统安全风险评方法研究"（项目编号：JAT160094）研究成果之一。
作者简介：张文德（ORCID：0000-0002-3017-9211），教授，博士生导师，E-mail：zhangwd@fzu.edu.cn；程涵（ORCID：0000-0003-2495-0898），硕士研究生；刘田，馆员；曾金晶，馆员。
收稿日期：2017-08-26 修回日期：2018-01-07 本文起止页码：119-124 本文责任编辑：王传清

教育数据挖掘算法的精度、效率和实现复杂度都有待提高,且对数据要求较高,在实际应用中存在一定阻力。

本文应用随机森林(Random Forest,RF)算法,充分利用其良好的泛化性和鲁棒性、对噪声不敏感、精度和准确性高的优点,构建基于随机森林的整合特征选择模型,并对高校贫困生认定数据进行实验分析,验证基于随机森林算法的高校信息整合特征选择模型的有效性和准确性,以期更好地进行高校信息整合,并提供个性化决策支持。

2 高校信息碎片化整合流程

2.1 高校信息碎片化整合

随着高校信息化建设的逐步实施,各大高校对于信息整合也做了很多有益的实践探索,如南京农业大学利用企业架构理论,搭建校园信息化应用架构平台,但由于大数据时代的到来,高校信息逐步趋于碎片化,而现有的信息整合体系存在动态可扩展性差、难以提供个性化决策支持等问题。同时,面对信息碎片化的影响,高校在学科建设、科研管理、学生管理等诸多方面存在不足,导致部门协同不够、目标指向各异、管理效率不高等一系列问题。因此,本研究定义一种"知识碎片",即通过对结构化、半结构化和非结构化的数据进行"碎片化"整合,得到的学校信息服务中最小颗粒度的知识片段[10]。当用户提出需求,我们只需分解该需求的关键特征,根据特征对"知识碎片"抽取和整合,形成可视化的查询结果并以报告的形式提交给用户。这种信息碎片化整合方式可以有效弥补现有高校整合系统可扩展性差、自主能力弱、信息利用率低的缺陷。

2.2 高校信息碎片化整合流程

针对信息碎片化的用户需求,本研究提出高校信息碎片化整合流程,该整合系统主要分两部分,一部分是信息整合过程,另一部分是用户访问过程。见图 1。

2.2.1 信息整合过程 信息整合过程主要针对高校众多异构数据库,如人事管理系统、教务管理系统、学工管理系统、科研管理系统、财务管理系统、资产管理系统以及一卡通系统等,获取这些系统中大量的结构化、半结构化和非结构化的数据,并通过统一的碎片化处理,将其转化成"知识碎片"的形式,并存储在知识碎片共享池中,实现高校资源的碎片化整合。

2.2.2 用户访问过程 用户访问过程主要根据用户提出的需求,查询历史需求特征库,判断是否存在相同需求的特征集合,若存在,则根据历史特征需求集合向知识碎片共享池中提取相应特征的"知识碎片",若不存在,则需要利用基于随机森林的特征选择模型提取满足需求的特征集合,再根据特征集合提取相应的"知识碎片",最终将符合需求特征的"知识碎片"加以整合,以可视化的形式反馈给用户。

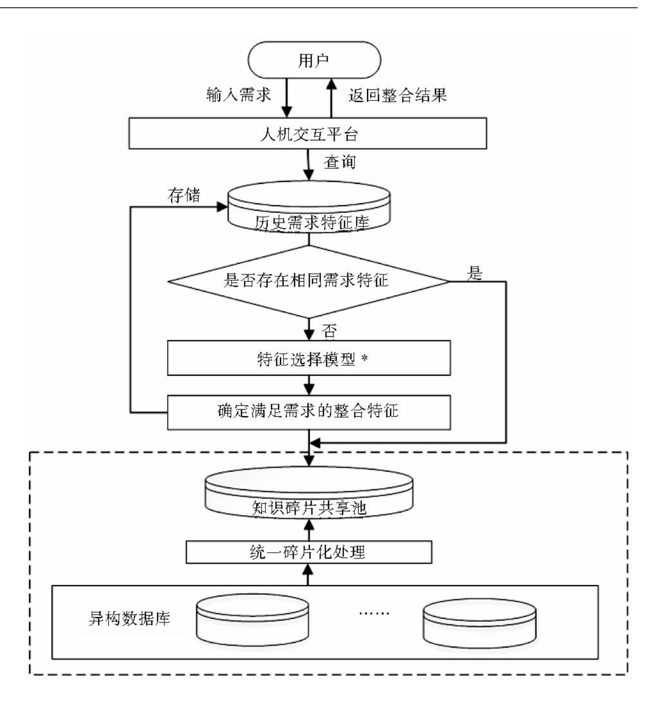


图 1 高校信息碎片化整合流程

因而,对于高校信息碎片化整合系统而言,其核心在于高校信息整合特征的选择过程,选择的特征准确性和有效性越高,则整合结果越具有可信度和说服力。而随机森林作为一种新型集成分类器,具有训练样本数量需求少、人工干预少、分类精度高等优点[11],可以处理高维数据并快速得到分类结果,可满足高校信息碎片化整合的需要。因此,高校信息碎片化整合系统可以有效应用到学科建设、科研管理和学生管理等工作中,实现"知识碎片"的集成及创新,从而为用户管理决策提供支持与帮助。

3 基于随机森林的高校信息碎片化整合特征选择模型构建

3.1 随机森林算法

随机森林是由美国学者 L. Breiman 于 2001 年提出的一个具有良好分类性能的机器学习算法,其基本思想是通过 Bagging 方法有放回的随机抽取不同的训练样本集,并对每个抽样样本构建相应的决策树,从而形成随机森林模型[12]。随机森林由一组决策树分类器

$\{h(x,\theta_k),k=1,2,\cdots,n\}$组成,其中每个决策树分类器中的$\theta_k$是独立同分布的随机变量,用来控制每一个决策树分类器的增长,变量n代表决策树分类器的数量,变量x代表输入的训练集样本数,每个决策树分类器根据输入的训练样本集产生分类结果,最终通过投票原则确定训练样本类别[13]。

随机森林算法的核心就是在训练过程中引入了随机性的思想。随机性的引入,可以降低各个决策树之间的相关度,从而提高随机森林的泛化性能,避免模型出现过拟合的现象。随机森林的随机性主要体现在两个方面:随机的训练样本子集和随机的特征子空间。随机的训练样本子集即采用 bagging 方法从样本中有放回地随机抽取n个与原样本集同样大小的训练样本集,这样可以保证初始训练集中约有63%的样本出现在样本集中。随机的特征子空间即在对决策树每个节点进行分裂时,从全部属性中等概率随机抽取一个属性子集,通常取值为$\sqrt{M}$个特征数,M为总特征个数,每次从这个子集中选择一个最优属性对当前节点中的剩余样本进行分裂[14]。

由于生成决策树的过程是独立的,因而随机森林算法可并行化处理,同时其分类精度高、训练速度快,不仅能有效克服过拟合问题,还能够评价特征重要程度[15－16]。因此,随机森林算法适用于高校信息碎片化整合过程中的特征选择。

3.2 特征的重要度计算

由于随机森林算法选择的特征是完全随机的,即每个特征被选中的概率完全相等,故认为每个特征对于目标需求的重要性相同。但在高校信息碎片化整合过程中发现大量的特征增加了模型的复杂度,并且对整合结果无明显影响,也就是说,实际上每个特征对于不同整合需求的重要度是不同的,对节点分裂影响也不同。因此需要在保证整合结果准确率的基础上,通过特征重要度计算,筛选出重要度较高的特征从而进行整合。

随机森林的特征重要性评分统计量计算有根据 Gini 指数和袋外数据(OOB)错误率两种方式[17－18]:,本研究根据 Gini 指数计算特征重要度。设一组随机变量$x_1,x_2,\cdots,x_M$,则变量x_j的得分统计量用$VIM_j^{(Gini)}$表示,其含义为第j个变量在随机森林的所有决策树中节点分裂不纯度的平均改变量。$VIM_j^{(Gini)}$的计算过程如下:

节点 的 Gini 指数为:

$$GI_m=\sum_{k=1}^{K}P_{mk}(1-P_{mk}) \qquad 公式(1)$$

其中K为样本集的类别数,P_{mk}为节点 m 样本属于第K类的概率估计值。

变量x_j在节点 m 的重要度为:

$$VIM_{jm}^{(Gini)}=GI_m-GI_l=GI_r \qquad 公式(2)$$

其中GI_l、GI_r分别表示由节点 m 分裂的两个新节点的 Gini 指数。

如果变量x_j在第i棵树中出现M次,则变量x_j在第i棵树的重要性为:

$$VIM_{ij}^{(Gini)}=\sum_{m=1}^{M}VIM_{jm}^{(Gini)} \qquad 公式(3)$$

变量x_j在随机森林中的 Gini 重要度为:

$$VIM_j^{(Gini)}=\frac{1}{n}\sum_{i=1}^{n}VIM_{ij}^{(Gini)} \qquad 公式(4)$$

其中n为随机森林中决策树的数量。

3.3 特征选择性能的评价指标

关于分类预测问题,常用的评价指标有查全率、查准率和分类精度等。针对本研究中的整合特征选择模型,同样定义这样一组评价指标[19]:算法的召回率(Rec)、算法的精确率(Pre)、算法分类结果的准确率(Acc)和 AUC(曲线下 roc 图的面积,area under roc curve)。

假设 TP 代表实际为正类且被确认为正类的个数,FP 代表实际为负类却被确认为正类的个数,FN 代表实际为正类却被确认为负类的个数,TN 代表实际为负类且被确认为负类的个数,则:

算法的召回率为:

$$\mathrm{Rre}=\frac{TP}{TP+FN} \qquad 公式(5)$$

表示正样本被正确分类占正样本的比例。

算法的精确率为:

$$\mathrm{Pre}=\frac{TP}{TP+FP} \qquad 公式(6)$$

表示正样本被正确分类占被分类为正样本的比例。

算法的分类准确率为:

$$\mathrm{Acc}=\frac{TP+TN}{TP+TN+FP+FN} \qquad 公式(7)$$

表示所有样本被正确分类的比例,该指标用来衡量总体分类的准确度,Acc 值越高则分类效果越好。

另外,由于高校信息整合所采用的样本数据通常正负比例不平衡,所以需要同时将 AUC 作为模型的评估指标之一[20－21],其计算公式为:

$$AUC = \frac{\sum rank - \frac{M(M+1)}{2}}{MN} \quad \text{公式(8)}$$

其中,M 为正类样本的数量;N 为负类样本的数量。

本研究使用 K 折交叉验证[22]的方法来估计分类精度,即将完整的数据集分成大致相等的 K 个子集,每次轮流使用不同的($K-1$)个子集训练模型,余下的一个子集测试模型,反复进行 K 次运算,最终将得到的评价指标均值作为该选择特征的指标估计值。

3.4 基于随机森林的整合特征选择模型设计

基于随机森林的高校信息碎片化整合特征选择模型,主要由 3 个模块构成:特征提取模块、训练模块和测试模块。如图 2 所示:

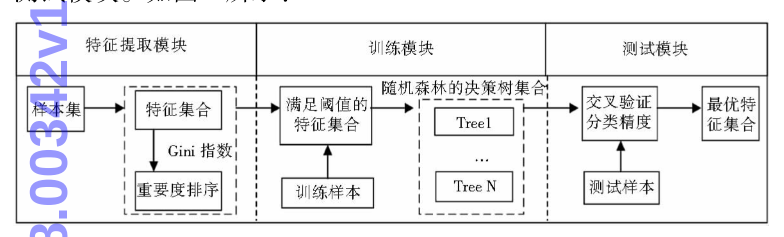


图 2 基于随机森林的整合特征选择模型

3.4.1 特征提取模块 该模块主要通过对样本集进行特征提取,形成所有特征的集合,再根据 Gini 指数计算所有特征的重要度,按从高到底的顺序对所有特征进行排序。由于大量特征向量不仅对整合结果没有影响,而且还提高了模型的复杂度,因此,需要设定一个阈值 λ,将排序结果与所选阈值进行比较,选取前 λ 个特征向量构成优化的特征集合进行训练。

3.4.2 训练模块 该模块主要将特征提取模块生成的优化特征集合输入训练模块,抽样一部分数据作为待训练样本,其余的作为测试样本。在训练样本上进行随机森林分类模型的创建,最终形成随机森林的决策树集合。

3.4.3 测试模块 该模块将测试样本输入训练后的决策树集,得出该特征集合的分类结果,并通过计算评价指标评价分类结果的精度,以此来确定特征选择的优劣。最终根据评价指标的最优结果确定整合特征的个数,并形成最优特征集合。

4 实验与结果讨论

4.1 实验背景及数据说明

我国高校贫困生认定过程存在很多困难,如无法判断在众多学生贫困指标中,哪些指标能反映学生贫困程度,如何控制人为主观因素的影响以及如何平衡贫困生认定过程的公开性和学生隐私的保密性问题等[23]。本实验以高校贫困生认定为例,验证随机森林在高校信息整合的特征选择过程中的准确性及有效性。

实验数据来源于某校某一年级 430 名学生,包括学生基本信息表、学生家庭情况表、学生消费情况表、学生贷款情况表、学生勤工助学表以及学生困难认定申请表等信息。数据集中正确判断学生为贫困生的样本数量记为 TP,正确判断学生不是贫困生的样本数量为 TN,学生本身为贫困生但判断不是贫困生的样本数量记为 FP,学生本身不是贫困生但判断是贫困生的样本数量记为 FN。

在实验过程中采用十折交叉验证的方法,即将数据集划分为训练集和测试集,其中训练集占总数据的 90%,用来设计和构造随机森林算法,其余 10% 作为测试集,用来检测算法的性能。

4.2 贫困生认定特征的重要度排序

利用随机森林模型构建 200 棵决策树进行贫困生认定的特征选择,选择相关特征并根据 Gini 指数进行重要度计算,取阈值 $\lambda = 27$,按重要度排序取前 27 个特征形成特征集合,其特征描述及重要度见表 1。

4.3 实验结果

大量文献表明,特征数量并非越多越好。特征数量过多不仅对整合结果没有影响,而且还提高了模型的复杂度。整合特征选择模型的目的在于通过样本数据的学习,找到满足用户需求的整合特征的最优集合,从而使得整合结果更加可靠。因此,我们需要通过对不同特征数量下的模型精度进行比较,并找到最优的整合特征集合。实验分别选取 3、6、9、12、15、18、21、24、27 个特征向量进行训练,计算指标评估结果如图 3 所示:

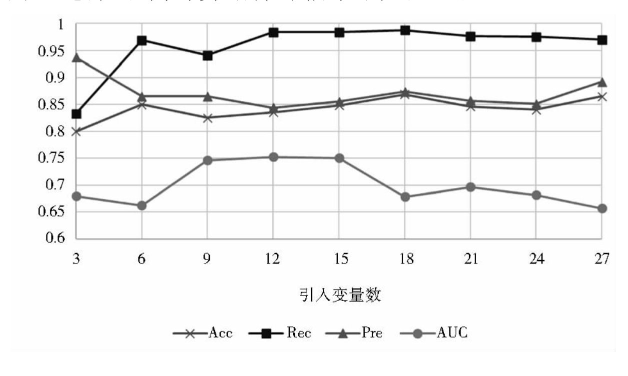


图 3 评估指标比较情况

表 1　贫困生评定特征重要度排序

特征序号	特征	特征描述	重要度
1	family_income	家庭月收入	1.0051E-01
2	isDisabled	是否为孤残	8.5436E-02
3	isHeavySick	有无重危病人	6.0030E-02
4	loan_amount	贷款金额	5.2865E-02
5	isLowIncome	是否为低保户	5.0087E-02
6	month_consum	平均月消费金额	4.9823E-02
7	isPoorPlace	是否来自贫困地区	4.8374E-02
8	isMakeUp	是否补考	4.7693E-02
9	haveDearConsume	有无奢侈消费	4.6351E-02
10	haveBigDisaster	是否遭遇重大灾害	4.5714E-02
11	family_number	家庭人口总数	4.2302E-02
12	numberInSchool	家庭在上学人数	4.0310E-02
13	IsApplyPoorStudent	是否申请贫困生	3.9964E-02
14	isWorking	是否勤工助学	2.8399E-02
15	monthWorkingIncome	月助学金额	2.8120E-02
16	major_ranking	专业成绩排名	2.7384E-02
17	health_condition	健康状况	2.7097E-02
18	isSingleParent	是否单亲	2.6133E-02
19	isMartyrChild	是否烈士子女	2.4649E-02
20	healthcondition_parents	父母健康状况	2.3208E-02
21	loan_timeLimit	贷款期限	2.1444E-02
22	course_credit	学分	2.0467E-02
23	class_job	担任职务	1.7513E-02
24	birthplace	生源地	1.7214E-02
25	working_class	助学种类	1.0617E-02
26	isfromCountry	是否农村户口	1.0609E-02
27	isActive	是否参与活动	7.6888E-03

根据图 3 可以得出以下结论：①特征向量个数小于 9 个时，精确率 Acc、召回率 Rec、准确率 Pre 及 AUC 都有较大幅度的变化，大于 9 个特征向量后，各指标开始趋于稳定，特征向量大于 15 个时，AUC 开始降低。通过对评估指标比较图中各评估指标在引入特征数量不同时的表现可知，并不是引入的特征向量越多，模型的拟合效果和预测结果就更好，因此，有必要对特征向量进行选择。②AUC 在 9－15 个特征向量时分类结果最优，AUC 高达 75%，Acc、Pre 都在 80% 以上，Rec 高达 95%。本次实验提取出 12 个最重要的特征变量，得出较为满意的评价结果，最优特征集合评价指标如表 2 所示：

表 2　最优特征集合的评价指标

准确率(Acc)/%	召回率(Rec)/%	精确率(Pre)/%	AUC
83.5443	98.4848	84.4156	75.1943

综上所述，针对高校贫困生认定问题，可根据家庭月收入、是否为孤残、有无重危病人、贷款金额、是否为低保户、平均月消费金额、是否来自贫困地区、是否补考、有无奢侈消费、是否遭遇重大灾害、家庭人口总数以及家庭在上学人数等 12 个特征进行信息碎片化整合，其准确率、召回率、精确率以及 AUC 都表现优异，说明该特征集合能够很好的为高校贫困生认定工作提供参考依据。

5　结语

本研究提出高校信息碎片化整合思想，即根据某一业务需求，选择满足需求的特征集合，根据最优的特征集合对知识碎片进行整合，从而构建了高校信息碎片化整合流程。高校信息碎片化整合流程的核心在于如何选择最优的整合特征，而随机森林良好的泛化性和鲁棒性、对噪声不敏感、能处理连续属性的特点，很适合用来构建高校信息整合特征选择模型。因此，本研究结合随机森林的优势，构建了基于随机森林的高校信息碎片化整合特征选择模型，并对模型的主要模块进行分析解读。

本研究通过高校贫困生认定这一实验验证了该模型在高校整合特征的选择上具有很高的准确性和精确度，为高校信息资源整合提供了一种可行的思路。随机森林虽然具有很好的辨识度，但该方法使得权值大的特征总是被选中，从而导致特征子空间的多样性降低，使得每棵决策树之间的相关性过高，反而使泛化误

差变大,因此后续需要对该算法进行改进,在提高特征相关性的同时,降低泛化误差。

参考文献:

[1] 马文峰,杜小勇. 基于数据的资源整合[J]. 情报资料工作, 2007(1): 41 -45.
[2] 马文峰,杜小勇,胡宁. 基于信息的资源整合[J]. 情报资料工作, 2007(1): 46 -50,70.
[3] 马文峰, 杜小勇,卢晓惠. 基于知识的资源整合[J]. 情报资料工作, 2007(1): 51 -56.
[4] 常桐善. 数据挖掘技术在美国院校研究中的应用[J]. 复旦教育论坛, 2009(2): 72 -79.
[5] 廖凤露,周庆. EDM 用于研究生就业能力的预测[J]. 教育教学论坛, 2017(33): 65 -66.
[6] 施佺,钱源,孙玲. 基于教育数据挖掘的网络学习过程监管研究[J]. 现代教育技术, 2016,26(6): 87 -93.
[7] 舒忠梅,徐晓东. 学习分析视域下的大学生满意度教育数据挖掘及分析[J]. 电化教育研究, 2014(5): 39 -44.
[8] 何世明,沈军. 基于 BP 神经网络的网上学习评价方法[J]. 微机发展, 2004,14(12): 26 -29.
[9] 刘美玲,李熹,李永胜. 数据挖掘技术在高校教学与管理中的应用[J]. 计算机工程与设计, 2010,31(5): 1130 -1133.
[10] 李恒贝, 查贵庭, 毛莉菊, 等. 基于碎片化服务的高校信息化架构及实践[J]. 中国教育信息化, 2016(19):11 -13.
[11] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3):32 -38.
[12] BREIMAN L. Random forests [J]. Machine learning, 2001, 45 (1): 5 -32.
[13] 周志华. 机器学习[M]. 北京: 清华大学出版社,2016: 178 -180.
[14] LIU Y, CHEN M. Random forest method and applicationin stream big data systems[J]. Journal of Northwestern Poly-technical University, 2015, 33(6): 1055 -1061.
[15] 吴辰文, 王伟, 李长生, 等. 一种结合随机森林和邻域粗糙集的特征选择方法[J]. 小型微型计算机系统, 2017, 38(6): 1358 -1362.
[16] YAO D, YANG J, ZHANG X. Feature selection algorithm based on random forest [J]. Journal of Jilin University (Engineering and Technology Edition), 2014,44(1):137 -141.
[17] 杨凯, 侯艳, 李康. 随机森林变量重要性评分及其研究进展[EB/OL]. [2017 -08 -25]. http://www. paper. edu. cn/html/releasepaper/2015/07/212/.
[18] ARCHER K J, KIMES R V. Empirical characterization of random forest variable importance measures [J]. Computational statistics & data analysis, 2008, 52(4): 2249 -2260.
[19] 王晓杰,孙仁诚,邵峰晶. 基于随机森林的用户对在线课程的放弃预测[J]. 青岛大学学报(工程技术版), 2016,31(4): 17 -22.
[20] 张晓凤,侯艳,李康, 基于 AUC 统计量的随机森林变量重要性评分的研究[J]. 中国卫生统计, 2016,33(3): 537 -540,542.
[21] JANUTZA S, STROBL C, BOULESTEIX A L. An AUC-based permutation variable importance measure for random forests [J]. BMC bioinformatics, 2013, 14(3): 433 -440.
[22] 王宇燕, 王杜娟, 王延章, 等. 改进随机森林的集成分类方法预测结直肠癌存活性[J]. 管理科学, 2017, 30(1):95 -106.
[23] 刘海苑. 基于数据挖掘的贫困生认定辅助系统的研究[J]. 电脑知识与技术, 2015, 11(24): 5 -7.

作者贡献说明:

张文德:提出研究思路,修订最终版本;
程涵:设计研究方案,起草论文;
刘田:设计实验方案及采集数据;
曾金晶:进行实验分析。

Application of Random Forest in the Fragmented Integration of University Information

Zhang Wende[1] Cheng Han[1] Liu Tian[2] Zeng Jinjing[2]
[1] Institute of Information Management, Fuzhou University, Fuzhou 350108
[2] Library of Fujian Agriculture and Forestry University, Fuzhou 350002

Abstract: [**Purpose/significance**] Facing the trend of fragmentation of university information, this paper puts forward the integration process of fragmented university information, and applies the random forest algorithm to construct the feature selection model of information-fragmented integration in universities. [**Method/process**] This paper represents the development, research status and existing problems of university information integration. Furthermore, in this paper, we elaborate the principles and advantages of the random forest algorithm, and use it to the feature selection model of information fragmented integration process in universities. Finally, we validate the model by using the example of identifying the students in the need of financial help. [**Result/conclusion**] Random forest algorithm shows higher accuracy and validity in the selection of features for integrating university information and therefore provides a new way for the integration of fragmented university information.

Keywords: random forest fragmentation information integration feature selection